Overview of the eHealth Knowledge Discovery Challenge at IberLEF 2020

Alejandro Piad-Morffis[a], Yoan Gutiérrez[b,c], Hian Cañizares-Diaz[a], Suilan Estévez-Velarde[a], Rafael Muñoz[b,c], Andrés Montoyo[b,c] and Yudivian Almeida-Cruz[a]

[a]School of Math and Computer Science, University of Havana, La Habana, 10400, Cuba
[b]Department of Language and Computing Systems, University of Alicante, Alicante, 03690, Spain
[b]University Institute for Computing Research, University of Alicante, Alicante, 03690, Spain

Abstract

This paper summarises the results of the third edition of the eHealth Knowledge Discovery (KD) challenge, hosted at the Iberian Language Evaluation Forum 2020. The eHealth-KD challenge proposes two computational tasks involving the identification of semantic entities and relations in natural language text, focusing on Spanish language health documents. In this edition, besides text extracted from medical sources, Wikipedia content was introduced into the corpus, and a novel transfer-learning evaluation scenario was designed that challenges participants to create systems that provide cross-domain generalisation. A total of eight teams participated with a variety of approaches including deep learning end-to-end systems as well as rule-based and knowledge-driven techniques. This paper analyses the most successful approaches and highlights the most interesting challenges for future research in this field.

Keywords

eHealth, Knowledge Discovery, Natural Language Processing, Machine Learning

1. Introduction

The vast amount of clinical text available online has motivated the development of automatic knowledge discovery systems that can analyse this data and discover relevant facts. These discoveries can be the base for novel treatments, understanding disease and drug interactions. Computational systems designed for this task are often trained on manually annotated corpora. To foster research in this area, the community has organised competitive challenges to identify, classify, extract, and link knowledge, such as in SEMEVAL [1] and CLEF campaigns [2].

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)
EMAIL: apiad@matcom.uh.cu (A. Piad-Morffis); ygutierrez@dlsi.ua.es (Y. Gutiérrez); hian.canizares@matcom.uh.cu (H. Cañizares-Diaz); sestevez@matcom.uh.cu (S. Estévez-Velarde); rafael@dlsi.ua.es (R. Muñoz); montoyo@dlsi.ua.es (A. Montoyo); yudy@matcom.uh.cu (Y. Almeida-Cruz)
ORCID: 0000-0001-9522-3239 (A. Piad-Morffis); 0000-0002-4052-7427 (Y. Gutiérrez); 0000-0002-5334-7468 (H. Cañizares-Diaz); 0000-0001-6707-1442 (S. Estévez-Velarde); 0000-0001-8127-9012 (R. Muñoz); 0000-0002-3076-0890 (A. Montoyo); 0000-0002-2345-1387 (Y. Almeida-Cruz)

[1] http://alt.qcri.org/semeval2020/
[2] http://www.clef-initiative.eu/

The eHealth Knowledge Discovery (eHealth-KD) challenge, in its third edition, leverages a semantic model of human language that encodes the most common expressions of factual knowledge, via a set of four general-purpose entity types and thirteen semantic relations among them. The challenge proposes the design of systems that can automatically annotate entities and relations in clinical text in the Spanish language. In this new edition, an alternative evaluation scenario (not related to the health domain) is also considered, which challenges participants to design systems that can successfully transfer their internal semantic representations from the health domain to an arbitrary new domain with considerably reduced training data. The challenge has been hosted at the Iberian Languages Evaluation Forum 2020, and included the participation of eight teams of researchers from different institutions.

This paper presents the design of the challenge as well as the data and tools provided to participants, and analyses the results obtained by each team. The remainder of the paper is organised as follows: Section 2 provides a detailed description of the tasks defined in the eHealth-KD challenge and the data provided for training and evaluation of knowledge discovery system, as well as all relevant evaluation metrics. Section 3 briefly describes all the solutions that were submitted to the challenge and introduces a set of characteristics that allow a qualitative comparison among them. Section 4 presents the main results of the challenge, divided into four evaluation scenarios, and analyses the most successful and promising approaches deployed by each team. Finally, Section 5 presents the conclusions of the research and recommendations for future editions.

2. Challenge description

The eHealth-KD challenge involves the identification of semantic entities and relations in natural language text. Although the focus has been on the health domain in past editions, the nature of the entities and relations extracted are general and can be applied to any domain. Figure 1 shows an example of three sentences with the relevant entities and relations annotated. An in-depth explanation of the annotation model is provided in Piad-Morffis et al. [1].

The evaluation of the challenge consists of submitting a set of natural language sentences with annotations automatically produced by a knowledge discovery system. Participants are provided with a set of manually annotated sentences (training and development corpus) that can be used for training and/or fine-tuning system as well as raw sentences that are used for evaluation (test corpus). The training and development corpus was provided two months in advance, but the test corpus was released only two weeks prior to the evaluation date, to discourage any fine-tuning on the test data. Although the actual source code of the system is not required, participants are encouraged to upload their code to open source code sharing services like Github.

To simplify the evaluation and provide more fine-grained comparisons the task is divided into two subtasks: one concerned with the identification and classification of entities, and the other concerned with the extraction of the semantic relations between these entities.

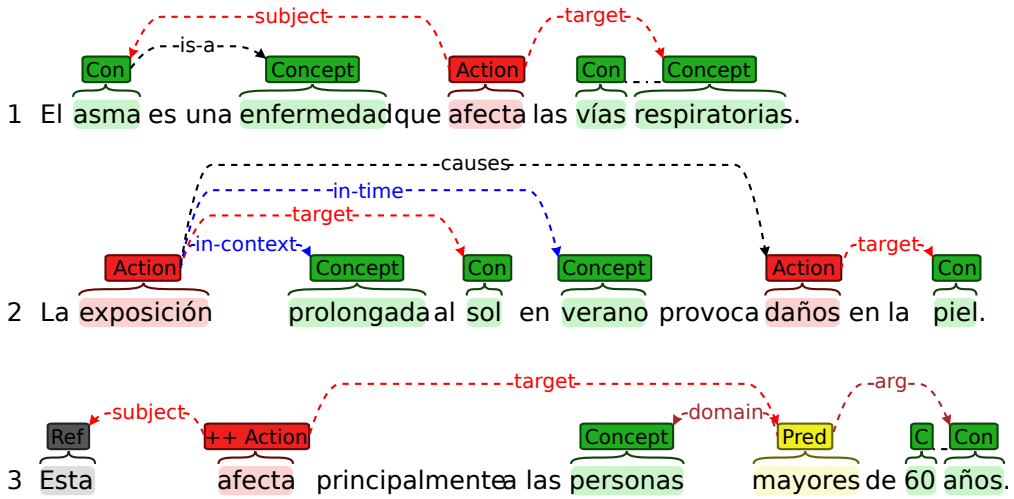

Figure 1: Example annotation of three sentences from the eHealth-KD challenge.

2.1. Subtask A: Entity Recognition

Given a list of eHealth documents written in Spanish, the goal of this subtask is to identify all the entities per document and their types. These entities are all the relevant terms (single word or multiple words) that represent semantically important elements in a sentence. The following figure shows the relevant entities that appear in a set of example sentences.

Some entities ("*vías respiratorias*" and "*60 años*") span more than one word. Entities will always consist of one or more complete words (i.e., not a prefix or a suffix of a word), and will never include any surrounding punctuation symbols, parenthesis, etc. There are four types for entities:

Concept: identifies a relevant term, concept, idea, in the knowledge domain of the sentence.

Action: identifies a process or modification of other entities. It can be indicated by a verb or verbal construction, such as "*afecta*" (affects), but also by nouns, such as "*exposición*" (exposition), where it denotes the act of being exposed to the Sun, and "*daños*" (damages), where it denotes the act of damaging the skin. It can also be used to indicate non-verbal functional relations, such as "*padre*" (parent), etc.

Predicate: identifies a function or filter of another set of elements, which has a semantic label in the text, such as "*mayores*" (older), and is applied to an entity, such as "*personas*" (people) with some additional arguments such as ''*60 años*" (60 years).

Reference: identifies a textual element that refers to an entity —of the same sentence or of different one—, which can be indicated by textual clues such as "*esta*", "*aquel*", etc.

2.2. Subtask B: Relation Extraction

Subtask B continues from the output of Subtask A, by linking the entities detected and labelled in the input document. The purpose of this subtask is to recognise all relevant semantic relationships between the entities recognised. Eight of the thirteen semantic relations defined for this challenge can be identified in Figure 1. The semantic relations are divided into the following categories:

General relations (6): general-purpose relations between two concepts (it involves Concept, Action, Predicate, and Reference) that have a specific semantic. When any of these relations apply, it is preferred over a domain relation –tagging a key phrase as a link between two information units–, since their semantic is independent of any textual label:

- **is-a:** indicates that one entity is a sub-type, instance, or member of the class identified by the other.
- **same-as:** indicates that two entities are semantically the same.
- **has-property:** indicates that one entity has a given property or characteristic.
- **part-of:** indicates that an entity is a constituent part of another.
- **causes:** indicates that one entity provokes the existence or occurrence of another.
- **entails:** indicates that the existence of one entity implies the existence or occurrence of another.

Contextual relations (3): enable an entity to be refined (it involves Concept, Action, Predicate, and Reference) by attaching modifiers. These are:

- **in-time:** to indicate that something exists, occurs or is confined to a time-frame, such as in "*exposición*" `in-time` "*verano*".
- **in-place:** to indicate that something exists, occurs or is confined to a place or location.
- **in-context:** to indicate a general context in which something happens, like a mode, manner, or state, such as "*exposición*" `in-context` "*prolongada*".

Action roles (2): indicate what role the entities play related to an Action:

- **subject:** indicates who performs the action, such as in "*[el] asma afecta [...]*".
- **target:** indicates who receives the effect of the action, such as in "*[...] afecta [las] vías respiratorias*". Actions can have several subjects and targets, in which case the semantic interpreted is that the union of the subjects performs the action over each of the targets.

Predicate roles (2): indicate which role play the entities related to a Predicate:

- **domain:** indicates the main entity on which the predicate applies.
- **arg:** indicates an additional entity that specifies a value for the predicate to make sense. The exact semantic of this argument depends on the semantic of the predicate label, such as in "*mayores [de] 60 años*", where the predicate label "*mayores*" indicates that "*60 años*" is a quantity, that restricts the minimum age for the predicate to be true.

2.3. Evaluation Scenarios

The eHealth-KD 2020 Challenge proposes four evaluation scenarios to measure different characteristics of the participant systems. We propose using a micro-averaged F_1 that weights all individual annotations equally, both entities and relations. Scenario 1 evaluates the solution to both tasks simultaneously, while Scenario 2 and 3 evaluate each task independently. Finally, Scenario 4 challenges systems to a novel domain with significantly less training data. This allows a more fine-grained comparison among systems with respect to specific capacities.

2.3.1. Main Evaluation (Scenario 1)

This scenario evaluates both subtasks together as a pipeline. The input consists only of a plain text, and the expected output is a BRAT `.ann` file with all the corresponding entities and relations found.

The measures will be precision, recall and F1 as follows:

$$Rec_{AB} = \frac{C_A + C_B + \frac{1}{2}P_A}{C_A + I_A + C_B + P_A + M_A + M_B}$$

$$Prec_{AB} = \frac{C_A + C_B + \frac{1}{2}P_A}{C_A + I_A + C_B + P_A + S_A + S_B}$$

$$F_{1AB} = 2 \cdot \frac{Prec_{AB} \cdot Rec_{AB}}{Prec_{AB} + Rec_{AB}}$$

The exact definition of Correct(C), Missing(M), Spurious(S), Partial(P) and Incorrect(I) is presented in the following sections for each subtask.

2.3.2. Optional Subtask A (Scenario 2)

This scenario only evaluates Subtask A. The input is a plain text with several sentences and the output is a BRAT .ann file with only entity annotations in it (relation annotations are ignored if present).

To compute the scores we define correct, partial, missing, incorrect and spurious matches. The expected and actual output files do not need to agree on the ID for each entity, nor on their order. The evaluation matches are based on the start and end of text spans and the corresponding type. A brief description about the metrics follows:

Correct matches are reported when a text in the development file —DEV— matches exactly with a corresponding text span in the gold file for START and END values, and also the entity type. Only one correct match per entry in the gold file can be matched. Hence, duplicated entries will count as Spurious.

Incorrect matches are reported when START and END values match, but not the type.

Partial matches are reported when two intervals [START, END] have a non-empty intersection, such as the case of "*vías respiratorias*" and "*respiratorias*" in the previous example (and matching LABEL). Notice that a partial phrase will only be matched against a single correct phrase. For example, "*tipo de cáncer*" could be a partial match for both "*tipo*" and "*cáncer*", but it is only counted once as a partial match with the word "*tipo*". The word "*cáncer*" is counted then as Missing. This aims to discourage a few large text spans that cover most of the document from getting a very high score.

Missing matches are those that appear in the GOLD file but not in the DEV file.

Spurious matches are those that appear in the DEV file but not in the gold file.

From these definitions, we compute precision, recall, and a standard F1 measure as follows:

$$Rec_A = \frac{C_A + \frac{1}{2}P_A}{C_A + I_A + P_A + M_A}$$

$$Prec_A = \frac{C_A + \frac{1}{2}P_A}{C_A + I_A + P_A + S_A}$$

$$F_{1A} = 2 \cdot \frac{Prec_A \cdot Rec_A}{Prec_A + Rec_A}$$

2.3.3. Optional Subtask B (Scenario 3)

This scenario only evaluates Subtask B. The input is plain text and a corresponding .ann file with the correct entities annotated. The expected output is a .ann file with both entities and relations. For this to happen, the entity annotations from the provided .ann file can be copied with the relation annotations appended.

To compute the scores we define correct, missing, and spurious matches. The expected and actual output files do not need to agree on the ID for each relation (which is ignored) nor on their order. The evaluation matches are based on the start and end of text spans and the corresponding type. A brief description about the metrics follows:

Correct: relationships that matched the GOLD file exactly, including the type and the corresponding IDs for each of the participants.

Missing: relationships that are in the GOLD file but not in the DEV file, either because the type is wrong, or because one of the IDs did not match.

Spurious: relationships that are in the DEV file but not in the gold file, either because the type is wrong, or because one of the IDs did not match.

We define standard precision, recall and F1 metrics as follows:

$$Rec_B = \frac{C_B}{C_B + M_B}$$

$$Prec_B = \frac{C_B}{C_B + S_B}$$

$$F_{1B} = 2 \cdot \frac{Prec_B \cdot Rec_B}{Prec_B + Rec_B}$$

2.3.4. Optional Alternative Domain Evaluation (Scenario 4)

This scenario evaluates a set of 100 sentences from an alternative domain (not health related), to experience with transfer learning techniques. A small development dataset with 100 sentences and their corresponding annotations will be provided when the general test set is released. Participants will need to train their systems in the full eHealth-KD 2020 corpus, and then apply some fine-tuning techniques in the additional 100 sentences from the alternative domain in order to successfully approach this scenario. The input and output format, and evaluation metrics are the same as for Scenario 1.

The purpose of this scenario, which we consider a complex challenge, is to stimulate the development of systems that can generalise to new knowledge domains without too many additional training examples. Hence, we encourage participants to focus not only on ehealth-specific features and techniques, but also consider more generalizable approaches.

2.4. Corpus Description

The corpus used in this edition of the challenge is composed of several sources reused from previous challenges, as well as new annotated content. The annotation guidelines and procedure followed were as described in Piad-Morffis et al. [2].

A total of 1000 training and development sentences are reused from the previous edition of the challenge, which is based on the same annotation model and methodology. For the test corpus, a new set of 300 sentences from Medline were manually annotated. An additional 200 sentences were selected from Wikinews, of which 100 were provided for development and 100 for testing in the evaluation Scenario 4. Finally, based on the submissions of the previous edition, an ensemble of 3,000 sentences automatically annotated was constructed by aggregating the annotations produced by previous participants. These sentences have not been manually revised, hence they are provided as an additional resource for fine-tuning but should be used with care when training a new system. The general statistics of the corpus are summarised in Table 1.

3. Systems Description

This section briefly describes the eight systems that were submitted to the challenge. In contrast with previous editions, there was a high degree of uniformity among participants, in the sense that most approaches involve the use of deep learning architectures with contextual or static embeddings. Nevertheless, there are interesting differences among the approaches which proved significant with respect to the results obtained. The participant teams and their corresponding systems are described next:

Table 1
Summary statistics of the eHealth-KD Corpus v2.0. Key phrases and relation labels are sorted by the number of instances in the training set. The training and development collections (marked with *) have been reused from previous editions.

Metric	Total	Training	DEV/Main	DEV/Transfer	Test	Ensemble
Sentences	3400	800*	200*	100	300	3000
Entities	25225	5012	1305	1242	2921	14745
- Concept	16207	3112	797	841	1944	9513
- Action	6431	1319	340	278	628	3866
- Predicate	1902	412	124	104	299	963
- Reference	685	169	44	19	50	403
Relations	20504	4571	1204	1241	2710	10778
- target	6376	1281	350	270	562	3913
- subject	3156	674	170	251	438	1623
- in-context	2503	502	140	193	380	1288
- is-a	2013	458	104	119	262	1070
- in-place	1250	304	77	111	237	521
- causes	890	292	71	30	92	405
- domain	994	269	74	82	196	373
- argument	857	254	73	47	185	298
- entails	308	117	43	11	28	109
- in-time	489	126	26	81	127	129
- has-property	1088	134	18	18	91	827
- same-as	346	93	31	19	66	137
- part-of	234	67	27	9	46	85

Vicomtech [3] presented an end-to-end deep neural network with pre-trained BERT models as the core for the semantic representation of the input texts. They experimented with two models: BERT-Base Multilingual Cased and BETO, a BERT model pre-trained on Spanish text. They model all the output variables—entities and relations—at the same time, modelling the whole problem jointly. Some of the outputs are fed back to the latter layer of the model, connecting the outcomes of the different sub-tasks in a pipeline fashion.

TALP-UPC [4] presented an end-to-end deep neural network, for simultaneously identifying key-phrases and their relationships, that does not rely on any domain-specific knowledge nor handcrafted features. Input documents are parsed using FreeLing and encoded using either a BERT, a Word2Vec or a FastText pre-trained word-embedding model. In order to generate all possible relations, the model should be run for every input token and have the all raw likelihoods combined across every one of them.

UH-MAJA-KD [5] presented a hybrid model for Subtask A that uses Stacked Bidirectional LSTM layers as contextual encoders, and linear chain Conditional Random Fields as tag decoders. The system addresses Subtask B in a pairwise query fashion, encoding information about the sentence and the given pair of entities using syntactic structures derived from the dependency parse tree, by the means of LSTM-based Recurrent Neural

Networks.

IXA-NER-RE [6] presented a two-step model for the NER and RE sub-task, each of them independently developed from the other. The Name Entity Recognition task has been envisaged as a basic seq2seq system applying a general-purpose Language Model and static embeddings. In the RE sub-task, two approaches were explored: transfer learning methods and Matching the Blank to tackle the problem of the reduced size of the training corpus by producing relation representations directly from unlabelled text.

UH-MatCom [7] presented several deep-learning models trained and ensembled to automatically extract the entities and relations. Their models use a combination of state of the art techniques such as BERT, Bi-LSTM, and CRF. They also explore the use of external knowledge sources such as ConceptNet.

SINAI [8] presented a BiLSTM+CRF neural network where different word embeddings are combined as an input to the architecture: custom-generated medical embeddings, contextualised non-medical embeddings, and pre-trained non-medical embeddings based on transformers.

HAPLAP [9] presented a joint AB-LSTM neural network which combines a Bi-LSTM with max pooling and an attentive Bi-LSTM for the relation extraction task. The Joint AB-LSTM is fed with the pre-processed sentences, their entities and relations between those, and distance embeddings.

ExSim [10] presented an information retrieval approach in which entities and relations in the training set are compared via word-embedding similarity to determine the most likely label.

Baseline is a basic implementation that stores all pairs of entities and labels, and all triplets of tow entities and relation labels found in the training set, and simply outputs for the test set a label if it finds an exact match. The purpose of the baseline is to provide participants with a starting point that already takes care of loading the data, parsing the annotation format, and producing the right output.

By far the most common type of approach corresponds to recurrent deep learning architectures (e.g., LSTM layers) with contextual embeddings (e.g., BERT). This combination is the basis of seven out of eight participant systems. This is not surprising given the recent success of these approaches in several NLP tasks, and in fact it was suggested in the Overview of previous editions of the eHealth-KD Challenge [11] [12]. Variations within this trend include the use of custom rather than pre-trained embeddings and the introduction of knowledge-based features. However, the most significant difference in approach corresponds to systems that perform an end-to-end strategy versus systems that solve each subtask separately. In the previous two editions of the challenge, the best performing system has used an end-to-end strategy. In this edition, two team (**Vicomtech** and **TALP**) deploy different end-to-end strategies.

3.1. Systems characteristics

For describing each system we define a set of characteristics that group the different approaches used by the participants. These characteristics span from abstract concepts as using external knowledge to implementation details such as using transformers or other contextual embeddings. The purpose of these characteristics is to analyse what is common among the systems that perform best in each scenario and possibly identify interesting or unexplored techniques. The characteristics are described below.

NLP: Using classic natural language processing features and strategies, such as TF-IDF encoding, stemming, lemmatization, dependency parsing, etc.

Static embeddings: Using pre-trained word embeddings such as Word2Vec or Glove, trained on standard corpora.

Contextual embeddings: Using contextual embeddings such as BERT or GPT, trained on standard corpora.

Custom embeddings: Using any type of embedding with a custom dataset selected for this task or a fine-tunning process.

Recurrent Network: Using any variant of recurrent neural networks, such as GRU or LSTM, possibly combined with other deep learning architectures.

Knowledge Bases: Using any source of external semantic knowledge either to define features or to enrich the training set.

End to end: Designing a single system that is simultaneously trained on both subtasks and shares at least a part of the features, representation or learning parameters for both entities and relations.

4. Results

Table 2 summarises the results obtained by each participant in each evaluation scenario. Results are sorted by F_1 in Scenario 1 which is considered the main evaluation. The top three results in each scenario are highlighted in bold.

Overall the best performing system was presented by **Vicomtech** [3] which not only obtains the best result in Scenario 1 (by a significant margin), but also ranks among the top three in all scenarios. Likewise, the system proposed by **Talp-UPC** [4] obtains the top result in Scenario 4, which is considered the most difficult scenario given the short number of training examples. It is also worth mentioning the results obtained by **UH-MAJA-KD**, who also rank among the top results in all scenarios, and the difference with the previous best result is less that 0.001 in two scenarios, which can be considered statistically insignificant.

Finally, it is interesting to note that the systems that obtained the best results for each individual task (i.e., **SINAI** in Scenario 2 and **IXA-NER-RE** in Scenario 3) do not rank among the top three in the general scenarios. This suggests an interesting trade-off between focusing on solving one specific task or designing a generally well-performing system.

Table 2
Results (F_1 metric) in each scenario, sorted by Scenario 1 (column *Score*). The top results per scenario are highlighted in **bold**.

Team	Score (F_1) Scn 1	Scn 2	Scn 3	Scn 4	Characteristics
Vicomtech	**0.665**	**0.820**	**0.583**	**0.563**	Recurrent Network, Contextual embedding, End-to-end
Talp-UPC	**0.626**	**0.815**	0.574	**0.583**	Recurrent Network, Contextual embedding, Static embedding, NLP, End-to-end
UH-MAJA-KD	**0.625**	0.814	**0.598**	**0.547**	Recurrent Network, Contextual embedding, NLP
IXA-NER-RE	0.557	0.691	**0.633**	0.478	Recurrent Network, Contextual embedding, Custom embedding
UH-MatCom	0.556	0.794	0.545	0.373	Recurrent Network, Contextual embedding, NLP, Knowledge Bases
SINAI	0.420	**0.825**	0.461	0.281	Recurrent Network, Contextual embedding, Custom embedding, Knowledge Bases
HAPLAP	0.395	0.541	0.316	0.137	Recurrent Network, Contextual embedding
ExSim	0.245	0.314	0.131	0.122	NLP, Static embedding

4.1. Analysis of Systems Performance

According to the characteristics defined in Section 3.1, we performed a qualitative analysis of the most successful strategies in each scenario. Figure 2 shows a box-plot of the ranking obtained by systems with each of the characteristics above defined, per evaluation scenario. The box-plot shows the mean, inter-quartile ranges, and the minimum and maximum score among all systems with a given characteristic.

As observed, the common strategy of using contextual embeddings and recurrent networks is capable of producing results in the full range of rankings. However, several systems have deployed and tailored this strategy, producing results with a range of variations. Hence, the use of BERT or LSTM layers alone does not guarantee a successful strategy. Likewise, as observed in previous editions, the use of custom embeddings seems to incur a marginal disadvantage, perhaps given that training high-quality embeddings in domain-specific corpora is difficult. On the other hand, the use of enternal knowledge bases to enrich semantic representations seems to be helpful in the entity recognition subtask, as exemplified by the result obtained by **SINAI** [8]. The single most successful approach seems to be the design of end-to-end architectures as opposed to solving both subtasks separately. This has been a trend in all the editions of the eHealth-KD challenge and is one of the most significant insights. The fact that end-to-end systems consistently outperform other approaches indicates that there is an interesting interaction between the semantic representation of entities and relations. Both end-to-end approaches presented provide an important advantage in terms of internal feedback exchange when resolving Subtask A and Subtask B, enhancing the discovery of entities and relations. This approach supports the idea that both subtasks are not completely independent of each other. However, as explained in Section 4, while end-to-end systems outperform all other approaches in Scenario 1 and 4, where both subtasks are performed, there are subtask-specific approaches that perform best when only one of the tasks is evaluated.

5. Conclusions and Future Work

The eHealth-KD 2020 proposed –as with the previous editions eHealth-KD 2019[11] and eHealth-KD 2018[12]– the modelling of human language in a scenario in which Spanish electronic health documents could be machine-readable from a semantic point of view. With this task, we

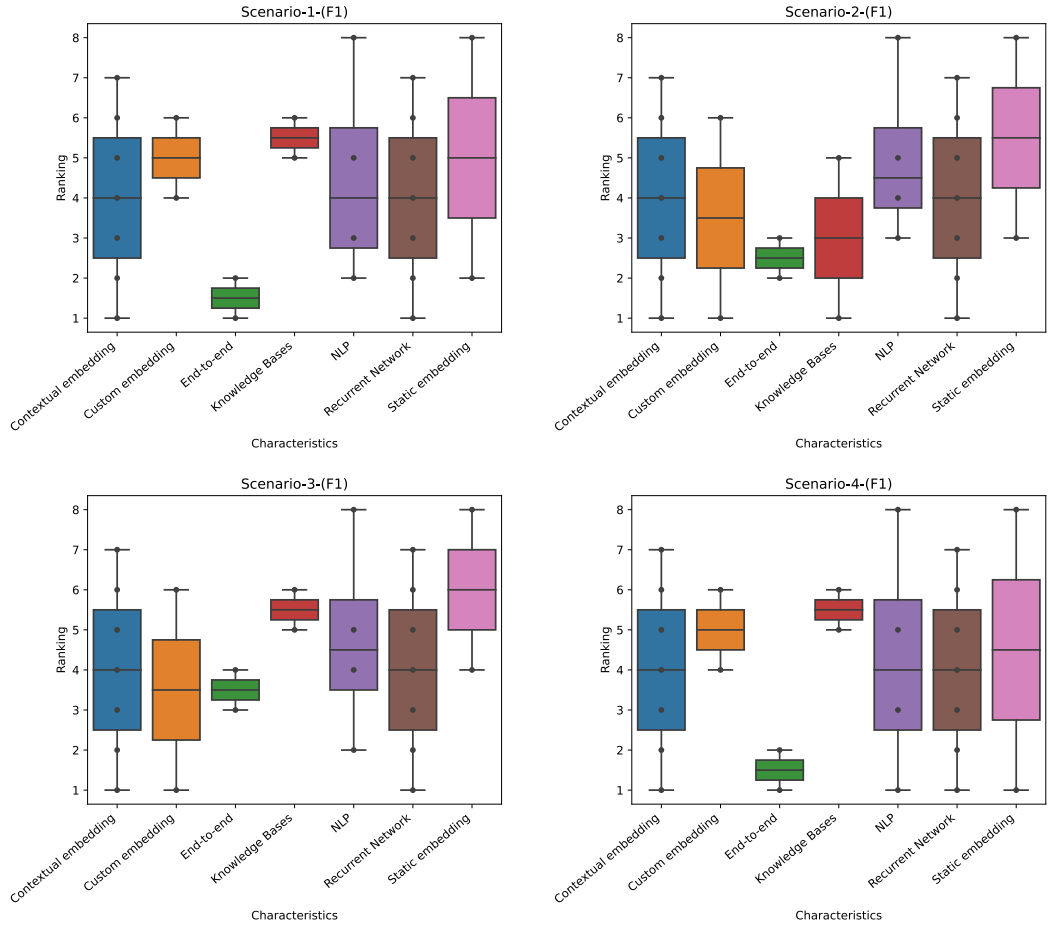

Figure 2: Box-plot of the distribution of ranking for the systems that applied each of the approaches defined in Section 3.1

.

expected to encourage the development of software technologies to automatically extract a large variety of knowledge from eHealth documents written in the Spanish Language. For this purpose, a new Spanish language corpus was manually annotated. Likewise, we provided tools to simplify the construction of knowledge discovery systems based on this corpus.

In the challenge, eight systems were presented achieving a maximum F1 score of 0.665. All participants presented algorithms in all scenarios, with the the end-to-end systems obtaining best results. The most used significant change to 2020's edition with respect to previous ones is the use of contextual embeddings (i.e., transformer architectures, and specifically BERT) as a replacement of static word embeddings. The results indicate that although promising approaches were presented in the challenge, the extraction of general-purpose semantic relations from natural language text is still an open area of research. Moreover, even though modern deep learning approaches are the most successful, we believe there is still a margin for improvement

by incorporating knowledge-based components that can exploit the structure of the annotation model.

Acknowledgments

This research has been partially supported by the University of Alicante and University of Havana, the Generalitat Valenciana (*Conselleria d'Educació, Investigació, Cultura i Esport*) and the Spanish Government through the projects SIIA (PROMETEO/2018/089, PROMETEU/2018/089) and LIVING-LANG (RTI2018-094653-B-C22).

References

[1] A. Piad-Morffis, Y. Guitérrez, S. Estevez-Velarde, R. Muñoz, A general-purpose annotation model for knowledge discovery: Case study in spanish clinical text, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 79–88.

[2] A. Piad-Morffis, Y. Gutiérrez, Y. Almeida-Cruz, R. Muñoz, A computational ecosystem to support ehealth knowledge discovery technologies in spanish, Journal of Biomedical Informatics (2020) 103517.

[3] A. García-Pablos, N. Perez, M. Cuadros, E. Zotova, Vicomtech at eHealth-KD Challenge 2020: Deep End-to-End Model for Entity and Relation Extraction in Medical Text, in: Proceedings of the Iberian Languages Evaluation Forum co-located with 36th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2020, 2020.

[4] S. Medina, J. Turmo, TALP at eHealth-KD Challenge 2020: Multi-Level Recurrent and Convolutional Neural Networks for Joint Classification of Key-Phrases and Relations , in: Proceedings of the Iberian Languages Evaluation Forum co-located with 36th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2020, 2020.

[5] A. Rodríguez Pérez, E. Quevedo Caballero, J. Mederos Alvarado, R. Cruz-Linares, J. P. Consuegra-Ayala, UH-MAJA-KD at eHealth-KD Challenge 2020: Deep Learning Models for Knowledge Discovery in Spanish eHealth Documents, in: Proceedings of the Iberian Languages Evaluation Forum co-located with 36th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2020, 2020.

[6] E. Andrés, O. Sainz, A. Atutxa, O. Lopez de Lacalle, IXA-NER-RE at eHealth-KD Challenge 2020: Cross-Lingual Transfer Learning for Medical Relation Extraction, in: Proceedings of the Iberian Languages Evaluation Forum co-located with 36th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2020, 2020.

[7] J. P. Consuegra-Ayala, M. Palomar, UH-MatCom at eHealth-KD Challenge 2020: Deep-Learning and Ensemble Models for Knowledge Discovery in Spanish Documents, in: Proceedings of the Iberian Languages Evaluation Forum co-located with 36th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2020, 2020.

[8] P. López-Ubeda, J. M. Perea-Ortega, D.-G. Manuel C., M. T. Martín-Valdivia, L. A. Ureña-López, SINAI at eHealth-KD Challenge 2020: Combining Word Embeddings for Named Entity Recognition in Spanish Medical Records, in: Proceedings of the Iberian Languages

Evaluation Forum co-located with 36th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2020, 2020.

[9] S. Santana, A. Pérez, A. Casillas, HapLap at eHealth-KD Challenge 2020, in: Proceedings of the Iberian Languages Evaluation Forum co-located with 36th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2020, 2020.

[10] Z. Hamzah Almugbel, ExSim at eHealth-KD Challenge 2020, in: Proceedings of the Iberian Languages Evaluation Forum co-located with 36th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2020, 2020.

[11] A. Piad-Morffis, Y. Gutiérrez, J. P. Consuegra-Ayala, S. Estevez-Velarde, Y. Almeida-Cruz, R. Muñoz, A. Montoyo, Overview of the ehealth knowledge discovery challenge at iberlef 2019, in: Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019, 2019, pp. 1–16. URL: http://ceur-ws.org/Vol-2421/eHealth-KD_overview.pdf.

[12] E. M. Cámara, Y. Almeida-Cruz, M. C. Díaz-Galiano, S. Estévez-Velarde, M. Á. G. Cumbreras, M. G. Vega, Y. Gutiérrez, A. Montejo-Ráez, A. Montoyo, R. Muñoz, A. Piad-Morffis, J. Villena-Román, Overview of TASS 2018: Opinions, health and emotions, in: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018, co-located with 34nd SEPLN Conference (SEPLN 2018), Sevilla, Spain, September 18th, 2018, 2018, pp. 13–27. URL: http://ceur-ws.org/Vol-2172/p0_overview_tass2018.pdf.